35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Classification of Alzheimer's Disease from FDG-PET images using Favourite Class Ensembles

Carlos Cabral[1], Margarida Silveira[2] and the Alzheimer's Disease Neuroimaging Initiative[3]

Abstract— **Classification of Alzheimer's disease (AD) and Mild Cognitive Impairment (MCI) from brain images using machine learning methods has become popular. Although the large majority of the existing techniques rely on a single classifier such as the Support Vector Machine (SVM), several ensemble methods such as Adaboost or Random Forests (RF) have also been explored. The ensemble methods combine the outputs of several classifiers and aim to increase performance by exploring the diversity of the base classifiers in terms of features or examples, which are usually randomly selected.**

In this paper we propose using a different kind of ensemble to address the three class problem of classifying AD, MCI and Control Normals (CN) from PET brain images. We propose the favourite class ensemble of classifiers where each base classifier in the ensemble uses a different feature subset which is optimized for a given class. Since different image features correspond to different sets of brain voxels, the proposed favourite class classifiers are able to take into account the fact that the spatial pattern of brain degeneration in AD changes in time as the disease progresses. We tested this approach on FDG-PET images from The Alzheimer's Disease Neuroimaging Initiative (ADNI) database using as base classifiers both Support Vector Machines (SVM) and Random Forests (RF). The ensembles systematically outperformed the corresponding single classifier with the best result (66.78%) being obtained by the SVM ensemble.

I. INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease that affects mostly elderly people. It is currently the leading cause of dementia worldwide and a growing cause of death in the developed countries. Mild Cognitive Impairment (MCI) is a transitional state from normal ageing (CN) to dementia and when it is associated with memory loss it is believed to be a precursor of AD [1]. There is no cure for AD but there are treatments that delay the progression of the disease and treat its symptoms. Naturally, the earliest the diagnosis is done, the most effective these therapeutics are [1].

Neuroimaging biomarkers such as Single-photon Emission Computed Tomography (SPECT), structural Magnetic Resonance Imaging (MRI) and Position Emission Tomography (PET) have been widely explored, where the later seems particularly suited for early detection since functional changes are proven to precede anatomical ones [1].

The machine learning methods used on these data mostly rely on a single classifier. The most widely used classifier is the Support Vector Machine (SVM), a powerful binary classifier, suited to high dimensional problems where few examples are available. It has been used, for example, in [2], [3] to classify MR images and in [4], [5] to classify PET images, all using voxel intensity (VI) as features. In a different approach a single multikernel SVM has been employed for the multimodal classification of MRI, PET and CSF using VI within regions of interest [6].

Although SVM has been the preferred single classifier, other options such as Fisher Linear discriminant [7], Gaussian Naives Bayes [8] or Gaussian Processes [9] have also been successfully used

The alternative to single classifiers is to use ensembles which combine the outputs of several classifiers and aim to increase performance by exploring the base classifiers diversity in terms of features or examples, which are usually randomly selected. Several well known ensemble methods have already been explored for AD classification. For instance [10] used Adaboost on the VI of PET images and [11] used RF on regions of interest from SPECT images. A different approach was suggested by [12] and applied to MRI images, where an ensemble classifiers was learned from different random subsets of local patches. Ensemble methods have also been used in order to combine information from different modalities such as EEG, MRI and PET [13].

Many of these methods use a prior feature selection step in order to reduce dimensionality. Different techniques have been used for this purpose, such as PCA [4] or selecting the best ranking features according to some criteria such as the t-test [6], Pearson correlation coefficient or Mutual Information [14].

In this paper we propose to classify AD, MCI and CN in PET brain images using an ensemble of classifiers approach, where base classifiers use different feature sets. However, the features used by the base classifiers are not fixed a priori nor randomly selected, they are class specific and they are selected as the ones that best discriminate each class from all the others. This is known as the favourite class method [15]. Using class specific features has advantages in terms of classification accuracy because features are not equally relevant for all classes but it has additional advantages, since it allows for the interpretation of the specific feature patterns associated with each class.

This work was supported by FCT project ADIAR PTDC/SAU-ENB/114606/2009.

[1] Carlos Cabral is with Institute for Systems and Robotics, Instituto Superior Técnico, Technical University of Lisbon, Avenida Rovisco Pais 1, 1049-001 Lisbon, Portugal cfcabral@isr.ist.utl.pt

[2] Margarida Silveira is with the Department of Electrical Engineering and Institute for Systems and Robotics, Instituto Superior Técnico, Technical University of Lisbon, Avenida Rovisco Pais 1, 1049-001 Lisbon, Portugal msilveira@isr.ist.utl.pt

[3] Data used in the preparation of this article were obtained from the ADNI database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

The results of the proposed method will be compared, for validation purposes, with the ones obtained by a classifier using features that are not class specific.

II. Material and Methods

A. Data

The data analyzed were retrieved from the ADNI database `http://www.adni-info.org/`. The ADNI initiative comprises a longitudinal multi-modal follow up of all participants across a period of time of 36 months in which imagiologic, clinical and biospecimen data were collected. Our study used rest-state Fluorodeoxyglucose (FDG) - PET brain volumes acquired 24 months after the first visit. Only subjects with Clinical Dementia Rating (CDR) of 0 for normal controls, 0.5 for MCI patients and 0.5 or higher for AD patients were chosen. After this selection process, we randomly selected a class balanced subset of 177 FDG-PET volumes. The FDG-PET images had previously been preprocessed by ADNI in four steps: co-registration, session average, standard space transformation and voxel intensity normalization and smoothing with a 8mm FWHM Gaussian filter. The resulting volumes were represented by a matrix of 128x128x60 yielding a total of 983040 voxels, of which only the ones corresponding to the brain were selected, in a total of 309881. The voxel intensity ranged from 0 to 32700. Table 1 presents a demographical (Age and Sex) and clinical (CDR and MMSE) characterization of the data used.

Group	CN	MCI	AD
Age(mean±)	77.4±4.9	77.7±6.9	78.2±6.6
Sex (M/F)	38/21	40/19	34/25
MMSE (mean± sd)	29.2±0.9	25.7±3	19.26±5.6
CDR (mean± sd)	0±0	0.5±0	1.2±0.6

TABLE I

Demographic and clinical characteristics of each group (Mean ± Standard deviation)

B. Feature Extraction and Selection

Our approach used the voxel intensities (VI) of each brain scan as the classification features. In order to select the subset of features used by the classifier, all the features were ranked according to their Mutual Information (MI) with the class label and the highest ranking features were then selected. Let $\mathbf{x_i} \in \Re^n$ denote the training patterns, $\mathbf{i}$ = 1,..., P and $\mathbf{y_i} \in \{1,2,3\}$ denote the corresponding classification. Mutual information is calculated as follows in Eq.1.

$$MI(X;Y) = \sum_{\chi \in X} \sum_{y \in Y} p(\chi,y) \log \frac{p(\chi,y)}{p(\chi) \cdot p(y)} \tag{1}$$

where densities were approximated using histograms, with as many bins as classes for Y and for X the number of bins is the cubic root of the number of examples in the training set.

This feature selection method was used in both the favourite class ensemble and in the base classifiers approach. However, there is a difference in the way it was applied in each case. In the favourite class ensemble each base classifier uses a different subset of the input features, which is the subset of features that best discriminates each class from all the others. Therefore, in this case the class label Y is a binary variable. To find the features subset of classifier $\mathbf{D_j}$, optimized for class $\omega_\mathbf{j}$, the class label took value 1 for patterns of that class and 0 for patterns of the other two classes. In the case of the base classifier approach, the subset of the input features used for classification is the one which jointly selects the most discriminate features for all three classes, thus the label Y is a ternary variable.

C. Classification

The base classifiers used in this study were SVM and RF. By using two different base classifiers we try to show that the proposed method is suited to different classifiers. The choice of base classifiers rested in RF and SVM as they are appropriate to high dimensional problems where relatively few training examples are available, and have been successfully used in neuroimaging problems. Both SVM and RF are discriminative classifiers as they try to approximate classification boundaries in the feature space instead of modelling the class-conditional density.

1) SVM: SVM is by definition a binary classifier that returns a class label. Therefore, in order to use SVM in the current ternary classification problem (AD, MCI and CN), a multi-class extension is required. We used LibSVM's [16] multi-class implementation which is based on combining the output of three pairwise classifiers. The method uses probabilistic outputs for the pairwise classifiers. These probabilistic outputs are obtained by Platt's method of mapping the decision values of SVM by means of a sigmoid function [17]. After the posterior probabilities provided by the pairwise classifiers have been obtained, they are combined in order to obtain the multi-class probability outputs. This is achieved by solving a linear system that minimizes the differences between the ratio of each pairwise classifier posterior probabilities and the ratio of the correspondent pair of final probability output [18]. Both Linear (L-SVM) and Gaussian Radial Basis Functions kernels (RBF-SVM) were used. Note that this three-class SVM base classifier is already an ensemble of classifiers by itself in which all classifiers in the ensemble are trained in the same feature space.

2) RF: RF is a version of "bagging" in which the base classifiers are random decisions trees. These modified decision trees are not deterministic as each node split is usually dependent on a small subset of features randomly selected. The main parameters of this classifier are the number of trees in the "forest" and the number of variables to use in the node decision split. Note that, as it happened with the SVM base classifier, the RF base classifier is also already an ensemble of decision trees.

3) Ensemble decision: Our ensemble approach creates three different favourite class ternary classifiers, each trained using a different feature subset which is optimized for a particular class. In order to combine their outputs several methods can be used [19]. Our best results were obtained using the mean combination rule which is a simple method and has been reported as one of the more accurate and stable non trainable combination methods [19]. For each test pattern **x**, the mean combination rule calculates the following support μ_j given to each class j=1, ..., L (L=3 in our case):

$$\mu_j(\mathbf{x}) = \frac{1}{L}\sum_{\mathbf{i}=1}^{L} d_{i,j}(\mathbf{x}) \qquad (2)$$

where $d_{i,j}$ represents the probability output of classifier $\mathbf{D_i}$ given to class $\omega_\mathbf{j}$. Finally, the pattern is classified in the class with maximum value of μ_j. A schematic diagram of the favourite class ensemble method is presented in Fig. 1.

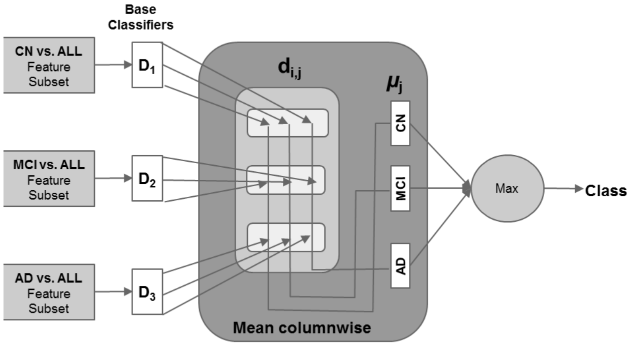

Base Classifiers
CN vs. ALL Feature Subset
MCI vs. ALL Feature Subset
AD vs. ALL Feature Subset
D1
D2
D3
d_{i,j}
μ_j
CN
MCI
AD
Max
Class
Mean columnwise

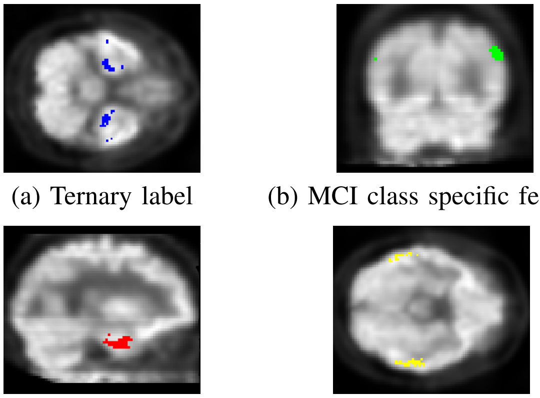
Fig. 1. Schematic diagram of the favourite class ensemble, adapted from [19]

III. Results

The above described methods were applied to the ternary problem of classifying between CN, MCI and AD. A number of VI features ranging from 5 to 10000 was used for both the single base classifiers and the corresponding favourite class ensembles. To evaluate the generalization performance of the method a 10-fold cross-validation procedure was used and the testing set accuracy was averaged across the folds. This cross validation experiment was repeated 5 times with fold randomization and the mean accuracy in all the experiments was taken. The SVM error tolerance parameter C used by the RBF and Linear kernels was estimated by nested cross validation in the training set within the range 2^{-10} and 2^5, following a geometric progression. The dispersion of the RBF kernel was fixed to the inverse of the number of features. In our Random Forest classification experiments we used Breiman's algorithm [20] as implemented in MatLab® with 100 trees and 2 features to split each node, based on results suggesting that performance was highly insensitive this parameter [20].

A. Pattern Analysis

Although there is a some variability of the spatial patterns we can make some general comments on the localization of the selected features for the single base classifier and the three sets selected in the favourite class ensemble method. Fig. 2 shows an example of the features selected for a representative number of features, 250. The method of feature selection used in the single base classifier approach has a strong preference for voxels located in the temporal and posterior cingulate cortex, whereas the three sets of features selected in the ensemble approach show some regional specialization and as a consequence greater variety. An interesting property of MCI specific features is their spatial distribution, they are highly asymmetric and clustered in the dorsolateral parietal cortex in conformity with previous studies [21]. It is pertinent to remark that the features chosen for the MCI category yield lower absolute values of MI. The maximum MI value for MCI is on average only 31,09% and 29,29% of the maxima for CN and AD respectively. This effect is probably a consequence of MCI definition as a transition state sharing clinical and neurological characteristics with both CN and AD, making the distinction between MCI and each the other two states harder.



(a) Ternary label (b) MCI class specific features

(c) CN class specific features (d) AD class specific features

Fig. 2. Anatomical localization of the features selected for a representative number of features, 250, for the three classifiers in the favourite class ensemble and the base classifier approach.

The pairwise overlap between the three feature sets used by the favourite class base classifiers was also analysed (plot not shown). In all cases it was under 3% in low dimensional spaces (less than 250 features). When a larger number of features were selected, the amount of overlap increased, as expected, but the increase was much sharper for the case of AD and CN overlap (maximum of 47.05%) than the overlap of MCI with the other two sets (maximum of 12.88%).

B. Classification Results

The results obtained for the classification experiments are shown in fig. 3, where it can be seen that in both cases the ensemble of classifiers performs better than the corresponding base classifier across almost all of the considered numbers of features. The ensemble superiority is most notable for SVM and for a small number of features, where there is less overlap between the features selected by the individual classifiers in the ensemble. We believe that the advantages of the favourite class ensembles rely on the fact that the class specific patterns can account for the distinct activity patterns that characterize each phase of the disease. Moreover, as the base classifiers return ternary probabilistic outputs, the favourite class ensemble is able to learn a profile

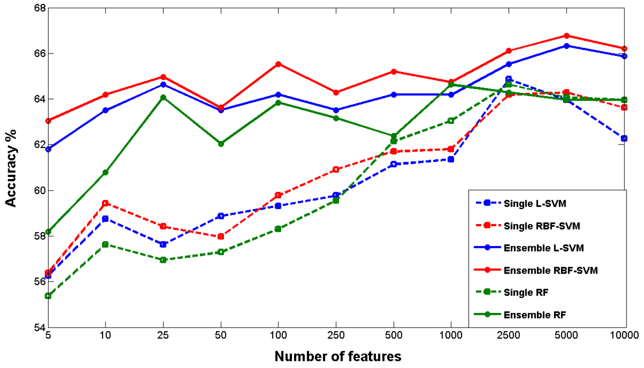


Fig. 3. Classification accuracy as a function of the number of features for the favourite class ensemble and single base classifiers for RF, L-SVM and RBF-SVM

for each class for all the class specific patterns. This way we are modelling the complex brain activity changes that characterize the transition from CN to AD in three stages, while the single base classifier approach takes into account the overall most informative features and aims to model this problem in a single step. In what concerns the results by base classifiers, L-SVM, RBF-SVM and RF, in the base classifier approach the performance is very similar for all, while in the favourite class ensembles RF performance is slightly poorer and the RBF-SVM is always superior.

In order to assess the statistical significance of the classification accuracy differences between methods, a variance analysis (ANOVA) was performed using as factors the classification method (favourite class ensemble or single base classifier), the base classifier (L-SVM, RBF-SVM or RF) and the number of features. Significant main effects were found for all the factors yielding p-values lower than 0.001. Also, significant interactions were found between the number of features and the classification method (p-value<0.001) and between the classification method and base classifier, although with a larger p-value, 0.0247.

IV. Conclusion

In this study we tested favourite class ensembles of RF and SVM classifiers to distinguish between FDG-PET brain images of CN, MCI and AD, in a ternary classification problem. Voxel intensity features were selected using Mutual Information as the filtering criterion. In our study the ensemble methods clearly outperformed the corresponding base classifiers across all the tested number of features for L-SVM and RBF-SVM base classifiers and almost all for RF as base classifier, with a peak accuracy of 66.78%, 66.33% and 64.63% for ensembles of RBF-SVM,L-SVM and RF, respectively.

The features selected are localized in the areas traditionally related with the progression of the AD in a compact and consistent pattern.

In the future we would like to develop ensembles with greater diversity of base classifiers, namely introducing other types of features and ensemble combination methods with learning capability.

References

[1] L. Minati, T. Edginton, M. G. Bruzzone, and G. Giaccone, "Current Concepts in Alzheimer's Disease: A Multidisciplinary Review," *American Journal of Alzheimer's Disease and Other Dementias*, Dec. 2008.

[2] Y. Fan, S. M. Resnick, X. Wu, and C. Davatzikos, "Structural and functional biomarkers of prodromal Alzheimer's disease: A high-dimensional pattern classification study," *NeuroImage*, vol. 41, no. 2, pp. 277–285, June 2008.

[3] N. Belmokhtar and N. Benamrane, "Article: Classification of Alzheimer's Disease from 3D Structural MRI Data," *International Journal of Computer Applications*, vol. 47, no. 3, pp. 40–44, June 2012, published by Foundation of Computer Science, New York, USA.

[4] I. Illán, J. Górriz, J. Ramírez, D. Salas-Gonzalez, M. López, F. Segovia, R. Chaves, M. Gómez-Rio, and C. Puntonet, "18F-FDG PET imaging analysis for computer aided Alzheimer's diagnosis," *Information Sciences*, vol. 181, no. 4, pp. 903–916, 2011.

[5] P. Padilla, M. López, J. Górriz, J. Ramírez, D. Salas-Gonzalez, and I. Álvarez, "NMF-SVM Based CAD Tool Applied to Functional Brain Images for the Diagnosis of Alzheimer's Disease," *Medical Imaging, IEEE Transactions on*, vol. 31, no. 2, pp. 207–216, 2012.

[6] D. Zhang, Y. Wang, L. Zhou, H. Yuan, and D. Shen, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.

[7] J. Stoeckel, G. Malandain, O. Migneco, P. M. Koulibaly, P. Robert, N. Ayache, and J. Darcourt, "Classification of SPECT Images of Normal Subjects versus Images of Alzheimer's Disease Patients," in *Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention*, ser. MICCAI '01. London, UK, UK: Springer-Verlag, 2001, pp. 666–674.

[8] M. Lopez, J. Ramirez, J. Gorriz, D. Salas-Gonzalez, I. Alvarez, F. Segovia, and R. Chaves, "Multivariate approaches for Alzheimer's disease diagnosis using Bayesian classifiers," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, 24 2009-nov. 1 2009, pp. 3190 –3193.

[9] J. Young, M. Modat, M. Cardoso, J. Ashburner, and S. Ourselin, "Classification of Alzheimer's disease patients and controls with Gaussian processes," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2012, pp. 1523 –1526.

[10] M. Silveira and J. Marques, "Boosting Alzheimer Disease Diagnosis using PET images," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, Aug. 2010, pp. 2556 –2559.

[11] R. Chaves, J. Ramírez, J. Górriz, I. Illán, F. Segovia, and A. Olivares, "Effective Diagnosis of Alzheimer's Disease by Means of Distance Metric Learning and Random Forest," *New Challenges on Bioinspired Applications*, pp. 59–67, 2011.

[12] M. Liu, D. Zhang, and D. Shen, "Ensemble sparse classification of Alzheimer's disease," *NeuroImage*, vol. 60, no. 2, pp. 1106 – 1116, 2012.

[13] R. Polikar, C. Tilley, B. Hillis, and C. Clark, "Multimodal EEG, MRI and PET data fusion for Alzheimer's disease diagnosis," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 31 2010-sept. 4 2010, pp. 6058 –6061.

[14] E. Bicacro, M. Silveira, J. Marques, and D. Costa, "3D brain image-based diagnosis of Alzheimer's disease: Bringing medical vision into feature selection," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2012, pp. 134 –137.

[15] N. Oza and K. Tumer, "Input decimation ensembles: Decorrelation through dimensionality reduction," *Multiple Classifier Systems*, pp. 238–247, 2001.

[16] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[17] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, 1999, pp. 61–74.

[18] T. Wu, C. Lin, and R. Weng, "Probability estimates for multi-class classification by pairwise coupling," *The Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[19] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] A. Venneri, "Imaging treatment effects in Alzheimer's disease," *Magnetic resonance imaging*, vol. 25, no. 6, pp. 953–968, 2007.